

CS657 Fall 2017

FINAL EXAM

NAME: Robert Jarvis

G#: G00647058

1. Given the following characteristic matrix for a set of 4 documents and 5 shingles,

Shingle	S1	S2	S3	S4
0	1	0	1	0
1	1	1	0	0
2	0	1	0	1
3	1	0	1	0
4	0	0	0	1

Answer the following questions

- a. Which pair of documents are the most similar (using Jaccard similarity)?
- b. Which documents contain shingle 0?
- c. Using the hashing functions:
 $h_1(n) = 3n + 2 \text{ mod } 7$, $h_2(n) = n - 1 \text{ mod } 7$, compute the matrix of signatures for the documents above. To what permutations do these hash functions correspond?
- d. Using the matrix of signatures and the algorithm covered in class, compute the signature matrix and compare the similarities found by using that matrix with the original similarities for the documents.

1) a) $\text{Jac}(s_1, s_2) = 1/4$

$$\boxed{\text{Jac}(s_1, s_3) = 2/3}$$

$$\text{Jac}(s_1, s_4) = 0/5$$

$$\text{Jac}(s_2, s_3) = 0/4$$

$$\text{Jac}(s_2, s_4) = 1/4$$

$$\text{Jac}(s_3, s_4) = 0/4$$

Jac s_1, s_3 most similar

b) s_1, s_3 contain simple zero

c) $h_1(n) \equiv 3n + 2 \pmod{7}$

$$h_2(n) \equiv n-1 \pmod{7}$$

$$\begin{aligned} h_1(0) &\equiv 2 \\ h_1(1) &\equiv 5 \\ h_1(2) &\equiv 1 \\ h_1(3) &\equiv 4 \\ \cancel{h_1(4)} &= 0 \end{aligned}$$

$$\begin{aligned} h_2(0) &\equiv 6 \\ h_2(1) &\equiv 0 \\ h_2(2) &\equiv 1 \\ h_2(3) &\equiv 2 \\ h_2(4) &\equiv 3 \end{aligned}$$

	s_1	s_2	s_3	s_4	
$h_1(0)$	2		2		
$h_2(0)$	6	1	6	1	
$h_1(1)$	2	5	2	1	
$h_2(1)$	0	0	6	1	
$h_1(2)$	2	1	2	4	
$h_2(2)$	0	0	6	1	
$h_1(3)$	2	1	2	1	
$h_2(3)$	0	0	2	1	
$h_1(4)$	2	1	2	0	
$h_2(4)$	0	0	2	1	

2	1	2	0
0	0	2	1

D) $\text{Jac}(s_1, s_2) = 1/2$

$$\text{Jac}(s_1, s_3) = 1/2$$

$$\text{Jac}(s_1, s_4) = 0$$

$$\text{Jac}(s_2, s_3) = 0$$

$$\text{Jac}(s_2, s_4) = 0$$

$$\text{Jac}(s_3, s_4) = 0$$

Figure 1 Question One

2. Starting with a (0.1,0.8,0.9,0.1) family, what are the results of amplifying it using the following constructs?
 - a. A 4-way AND followed by a 5-way OR
 - b. A 5-way OR followed by a 4-way AND
 - c. Which of the two has a large False Positive rate?

Q2

$$2) \text{ AND} \rightarrow \text{OR } (1 - (1-p)^b)^r \quad \text{OR} \rightarrow \text{AND } (1 - (1-p)^b)^r$$

$$\text{Family} = (0.1, 0.8, 0.9, 0.1)$$

$$\text{a) } r=4 \quad b=5 \quad (1 - (1-p)^b)^r$$

P	$1 - (1-p)^b$	$(0.1, 0.8, 0.99518, 0.00049)$
0.1	0.00049	
0.2	0.00797	
0.3	0.03964	
0.4	0.12161	
0.5	0.27580	
0.6	0.50043	
0.7	0.74661	
0.8	0.92826	
0.9	0.99518	

P	$(1 - (1-p)^b)^r$	$r=4 \quad b=5$
0.1	0.02812	
0.2	0.2043	
0.3	0.4790	
0.4	0.7233	
0.5	0.8807	
0.6	0.9596	
0.7	0.9903	
0.8	0.9987	
0.9	0.9999	

c) 5-way OR followed by 4-way AND has larger false positive rate

Figure 2 Question Two

3. Compute sketches of the vectors $[2,3,4,5]$, $[-2,3,-4,5]$ and $[2,-3,4,-5]$ using the following random vectors $v_1 = [1,1,1, -1]$, $v_2 = [1,1, -1, -1]$, $v_3 = [1,1,1,1]$, $v_4 = [-1,1,1,1]$. Compute the true and estimated cosines between the vectors .

Question 3

v_1	$\begin{bmatrix} 2, 3, 4, 5 \end{bmatrix}$			
v_2	$\begin{bmatrix} -2, 3, -1, 5 \end{bmatrix}$			
v_3	$\begin{bmatrix} 2, -3, 4, -5 \end{bmatrix}$			
Rv_1	$\begin{bmatrix} 1, 1, 1, -1 \end{bmatrix}$			
Rv_2	$\begin{bmatrix} 1, 1, -1, -1 \end{bmatrix}$			
Rv_3	$\begin{bmatrix} 1, 1, 1, 1 \end{bmatrix}$			
Rv_4	$\begin{bmatrix} -1, 1, 1, 1 \end{bmatrix}$			

SKETCHES

Sketches

$v_1 = \begin{bmatrix} 1, -1, 1, 1 \end{bmatrix}$

$v_2 = \begin{bmatrix} -1, -1, 1, 1 \end{bmatrix}$

$v_3 = \begin{bmatrix} 1, 1, -1, -1 \end{bmatrix}$

Random

estimate	cosine	Radian	Degree
$v_1, v_2 = 1.25 \cdot 180$	$= 45 = \cos(45) = 0.7071$	0.7071	
$v_1, v_3 = 1.5 \cdot 180$	$= 90 \quad \cos(90) = 0.0$	0.0	
$v_2, v_3 = 1.75 \cdot 180$	$= 135 \quad \cos(135) = -0.7071$	-0.7071	

TRUE COSINE

v_1, v_2 : angle = $(2 \cdot 2) + (3 \cdot 3) + (4 \cdot 4) + (5 \cdot 5)$

$$\frac{-4 + 9 + 16 + 25}{-4 + 9 + 16 + 25} = \sqrt{54}$$

v_2, v_3 : angle = $(-2 \cdot 2) + (3 \cdot 3) + (-4 \cdot 4) + (5 \cdot 5)$

$$\frac{-4 + 9 + 16 + 25}{-4 + 9 + 16 + 25} = \sqrt{54}$$

$$= \frac{14}{\sqrt{54}} = 0.259$$

v_1, v_3 angle = $(2 \cdot 2) + (3 \cdot -3) + (4 \cdot 4) + (5 \cdot -5)$

$$\frac{4 + -9 + 16 + -25}{4 + -9 + 16 + -25} = \frac{-5 + 16 + -25}{-5 + 16 + -25} = \frac{-30 + 16}{-30 + 16} = -14$$

$v_1 = \sqrt{2^2 + 3^2 + 4^2 + 5^2} = \sqrt{54}$

$v_3 = \sqrt{2^2 + 3^2 + 4^2 + 5^2} = \sqrt{54}$

-0.259

v_2, v_3 angle = $(-2 \cdot 2) + (3 \cdot -3) + (-4 \cdot 4) + (5 \cdot -5)$

$$\frac{-4 + -9 + -16 + -25}{-4 + -9 + -16 + -25} = \frac{-13 + -16 + -25}{-13 + -16 + -25} = \frac{-54}{-54} = 1$$

Figure 3 Question Three

4. Consider the following collection of baskets

{1, 2, 3} {2, 3, 4} {3, 4, 5} {4, 5, 6}
{1, 3, 5} {2, 4, 6} {1, 3, 4} {2, 4, 5}
{3, 5, 6} {1, 2, 4} {2, 3, 5} {3, 4, 6}

And a support threshold of 4. We are using PCY and a hash table of 11 buckets where the pair (i,j) is hashed to $ixj \bmod 11$

In the second stage, we hash pairs to 9 buckets using $i+j \bmod 9$.

Answer the following questions:

- a. Which buckets in the first pass are frequent?
- b. Which pairs are counted on the second pass?
- c. What are the counts of the buckets on the second pass?
- d. Does the second pass reduce the number of candidates?

Question 4

$(1,2)$	$(1,3)$	$(2,3)$	$(1,5)$
2	3	3	1

$(3,5)$	$(3,6)$	$(2,4)$	$(2,6)$
4	2	4	1

$(1,4)$	$(3,4)$	$(2,5)$	$(4,5)$
2	4	2	3

$(4,6)$	$(5,6)$	$\text{Support} = 4$
3	2	

	1	2	3	4	5	6	7	8	9	10
0	$(2,6)$ $(3,4)$	$(1,2)$ $(4,6)$	$(1,3)$ $(3,5)$ $(1,4)$	$(3,5)$ $(1,4)$	$(1,5)$	$(2,3)$ $(2,6)$	$(2,4)$ $(3,5)$ $(5,6)$	$(4,6)$ $(2,4)$ $(3,5)$	$(4,5)$ $(3,5)$	$(3,5)$

a) Frequent 1, 2, 4, 8 b) $(2,6), (3,4), (1,2)(4,6)$
 $(3,5), (1,4), (2,4), (5,6)$

c) counts

	1	2	3	4	5	6	7	8
0	$(4,6)$	$(5,6)$	$(1,2)$		$(1,4)$	$(2,4)$	$(3,4)$	$(2,6)$ $(3,5)$

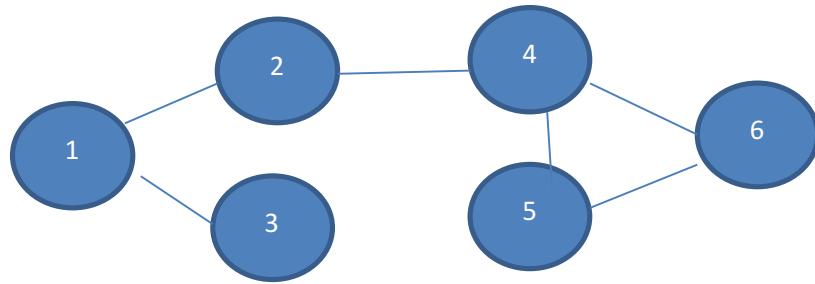
Frequent = 6, 7, 8

D) Reduces buckets by one and

of pairs from 8 to 4

Figure 4 Question Four

5. Consider the graph



Compute the Laplacian matrix, find its first two eigenvectors and suggest two different partitions of the graph based on those eigenvectors.

Question 5

Laplacian Matrix

$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & -1 & -1 & 0 & 0 & 0 \\ 3 & -1 & 2 & 0 & -1 & 0 \\ 4 & 0 & 1 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & -1 & 2 \\ 6 & 0 & 0 & 0 & -1 & -1 \end{bmatrix}$$

$$\begin{bmatrix} 2-\lambda & 1 & -1 & 0 & 0 & 0 \\ 1 & 2-\lambda & 0 & -1 & 0 & 0 \\ -1 & 0 & 2-\lambda & 0 & 0 & 0 \\ 0 & -1 & 0 & 3-\lambda & -1 & 0 \\ 0 & 0 & 0 & -1 & 2-\lambda & 0 \\ 0 & 0 & 0 & 0 & -1 & 2-\lambda \end{bmatrix}$$

λ_1 λ_2 λ_3

LAPLACIAN MATRIX

Eigenvalue	0	3	3	
Eigen vector	1	2	2	N_1
	1	-1	-1	N_2
	1	-1	-1	N_3
	1	-1	-1	N_4
	1	0	1	N_5
	1	1	0	N_6

Partition ONE

threshold zero on λ_2

Group 1 = { N_1, N_5, N_6 }

Group 2 = { N_2, N_3, N_4 }

split the negative values from

the positive values on λ_2

Partition two

Group 1 { N_5, N_6 }

Group 2 { N_1, N_2, N_3, N_4 }

split values across λ_1 and λ_2

keep 0's and 1's together

